Supplementary Material

Patchwork: A Patch-wise Attention Network for Efficient Object Detection and Segmentation in Video Streams

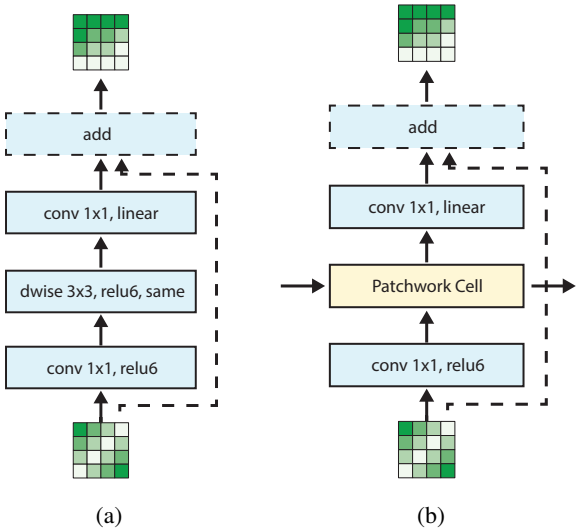
A. Detailed network architectures

Here, we discuss in details the network architecture. All Patchwork experiments use the MobileNetV2 [33] backbone, which consists of bottleneck blocks as shown in Fig. 9a. One bottleneck block consists of a 1x1 convolution to expand the number of channels, a 3x3 separable convolution layer with an optional stride, and a 1x1 convolution to project the number of channels back. Fig. 9b shows the modified MobileNetV2 block, where we replace the separable convolution with the *SAME* padding with a stateful Patchwork Cell that contains a separable convolution with the *VALID* padding.

We concatenate such stateful blocks together to form our stateful MobileNetV2 backbone network. An exact description appears in in Tab. 3.

For the object detection task in Sec. 4.1, the detector head builds on top of the modified stateful MobileNetV2, as shown in Fig. 10. Smaller objects are predicted by the high-resolution SSD heads that operate directly on the partial feature map. The memory of a Patchwork Cell restores a feature map for the full-sized frame, which allows the prediction of large objects within the SSD [25] framework. The feature pyramid consists of 4 layers of separable convolutions, which gradually reduce the spatial resolution. The attention network, which consists of two convolution layers and one fully-connected layer also builds on top of the restored full-sized feature map.

The segmenter head in Sec. 4.2 builds on top of the modified stateful MobileNetV2 backbone. Since the pixel-wise segmentation operates locally, it does not require a restored full-frame-sized feature map.



(a) (b)

Figure 9: a): A regular MobileNetV2 block [33]. The skip connection and the "add" op (dotted) are present only when the stride is 1. b): A modified MobileNetV2 block that includes a stateful Patchwork Cell as described in Sec. 3.2.

Input	Operator	t	c	n	s
$96^2 \times 3$	s-conv2d	-	32	1	2
$48^2 \times 32$	s-bottleneck	1	16	1	1
$48^2 \times 16$	s-bottleneck	6	24	2	2
$24^2 \times 16$	s-bottleneck	6	32	3	2
$12^2 \times 32$	s-bottleneck	6	64	4	2
$6^2 \times 64$	s-bottleneck	6	96	3	1
$6^2 \times 96$	s-bottleneck	6	160	3	2
$3^2 \times 160$	s-bottleneck	6	320	1	1
$3^2 \times 320$	s-conv2d	-	1280	1	1
$3^2 \times 1280$	-	-	-	-	-

Table 3: Our stateful MobileNetV2 backbone [33] for the M=4,N=1 Patchwork configuration. **s-conv2** is a stateful convolution layer with the Patchwork Cell, as shown in Fig. 1b. **s-bottleneck** is the stateful MobileNetV2 block with the Patchwork Cell, as shown in Fig. 9a. t is the MobileNetV2 expansion factor as described in [33]. c is the output channels, n the number of repetition for the layer and s is the stride for the first layer of a kind, repeated layers have the stride 1. The values in this table for the M=2,N=1 and M=4,N=2 setups remain the same, except for doubled height and width in all layers.

B. Patchwork Cell approximation

The Patchwork Cell in Sec. 3.2 approximates lost context using features from the past. The most accurate way to do so is to apply a single Patchwork Cell at the start of the network to recover as much context as possible, as shown in

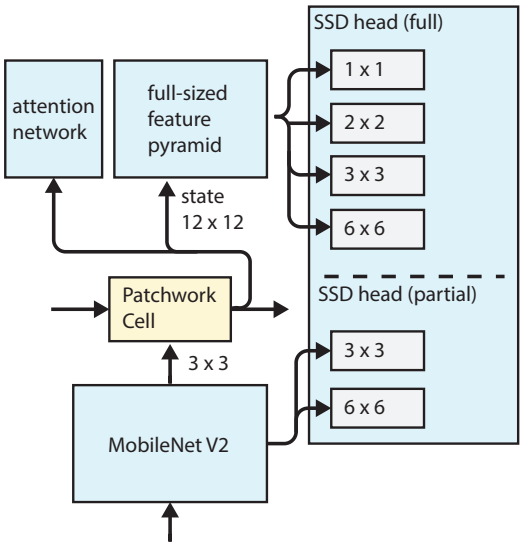


Figure 10: A detailed look at the Patchwork object detector head. High-resolution SSD [25] heads operate on the cropped feature maps, and only produces predictions for the partial window. A Patchwork cell restores a full feature map from its memory, on top of which we build a feature pyramid. Low-resolution SSD heads then operate on this feature pyramid.

Fig. 11a. However, doing this nullifies most of the latency saving that is the motivation of Patchwork. Alternatively, a series of Patchwork Cells can approximate the context incrementally (see Fig. 11b), which preserves the latency saving while suffers less than 0.1% of accuracy loss.

C. Detailed results

Tab. 4 and Tab. 5 show a more detailed view of the quantitative results in Fig. 5, Fig. 6, Fig. 7, Fig. 8. We report the average latency in addition to the maximum latency in both the empirical and theoretical measure. For the empirical latency, we also show the mean and standard deviation. Notably, the standard deviation for any baseline with an interval larger than 1 is high due to the uneven nature of the latency that only occurs every K keyframes. The high variance might be undesirable for many applications. For the maximum empirical latency, we show both the 95 and the 99 percentile.

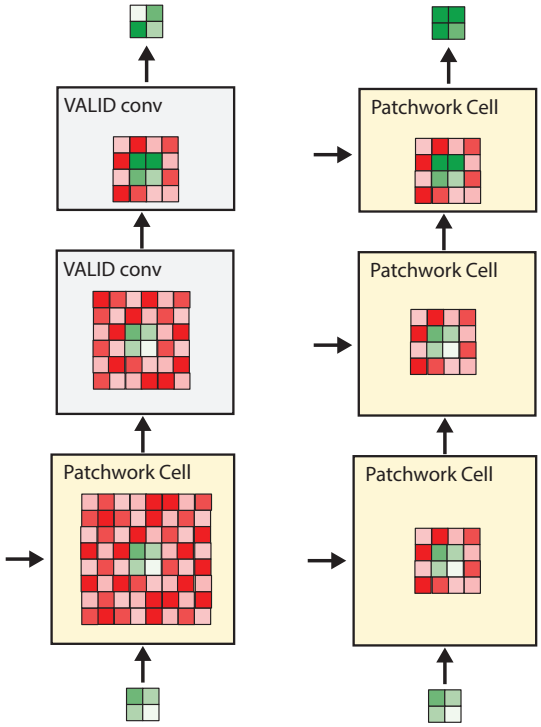


Figure 11: Comparison between different Patchwork Cell usage. Green indicates features extracted from the crop at the current time step. Red indicates features from the past provided by the Patchwork Cell. a) An architecture where the context is approximated by one single Patchwork Cell in the first layer. b): The proposed architecture that gradually approximates the lost context via Patchwork Cells throughout the network.

ID	Method	MFLOPs		msecs		MAP ↑
		max	avg	max	avg	
1	Single-frame (SF)	2047	2047	152.7 / 160.6	134.3±8.9	54.7
2	SF interval=4	2047	512	146.9 / 155.0	33.8±58.8	53.4
3	SF interval=16	2047	128	127.6 / 147.8	8.41±32.7	45.7
4	Patchwork M=4, N=2, depth=1.4, flip	1945	1945	149.2 / 156.3	137.7±6.2	**58.7**
5	Patchwork M=4, N=2, depth=1.4	973	973	75.2 / 79.2	68.1±3.6	57.4
6	SF depth=0.5	602	602	73.4 / 80.9	66.3±3.7	47.2
7	SF resolution=0.25	512	512	47.4 / 50.0	41.7±2.6	46.8
8	SF interval=4, delay=3	512	128	36.7 / 38.8	8.5±16.9	49.3
9	Patchwork M=4, N=2	543	543	55.2 / 58.4	47.9±3.7	**54.3**
10	SF interval=16, delay=15	128	8	8.0 / 9.2	0.5±2.0	34.3
11	Patchwork M=4, N=1	162	162	28.1 / 29.8	24.2±2.3	**41.6**

Table 4: A latency vs. accuracy comparison between the single-frame and Patchwork variants ImageNet VID. For the empirical (msecs) metric, we report the 95 and 99 percentile latency for **max** and the mean/std pair for **avg**. Methods with similar maximum FLOPs are grouped together. Same notation as described in Sec. 4.

ID	Method	FLOPs		msecs		$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
		max	avg	max	avg		
1	Single-frame (SF)	3307	3307	190.5 / 205.0	174.0±9.2	63.6	59.5
2	SF interval=4	3307	827	171.5 / 181.7	41.0±71.3	55.8	47.9
3	SF interval=16	3307	207	156.5 / 174.3	10.3±40.0	40.5	31.7
4	SF depth=0.5	1240	1240	108.6 / 114.8	99.5±5.3	58.4	55.2
5	SF resolution=0.25	827	827	51.6 / 58.5	45.2±3.5	56.6	50.8
6	SF interval=4, delay=3	827	207	42.9 / 45.4	10.3±17.8	44.6	34.1
7	Patchwork M=4, N=2	841	841	65.5 / 71.1	58.5±3.6	**62.1**	**58.8**
8	SF interval=16, delay=15	207	52	9.8 / 10.9	0.64±2.5	27.5	20.8
9	Patchwork M=4, N=1	221	221	33.4 / 36.2	28.3±2.7	**42.2**	**36.9**

Table 5: Object segmentation results on DAVIS 2016. The experimental setups and markings are similar to those in the detection results, see Tab. 4 for details.